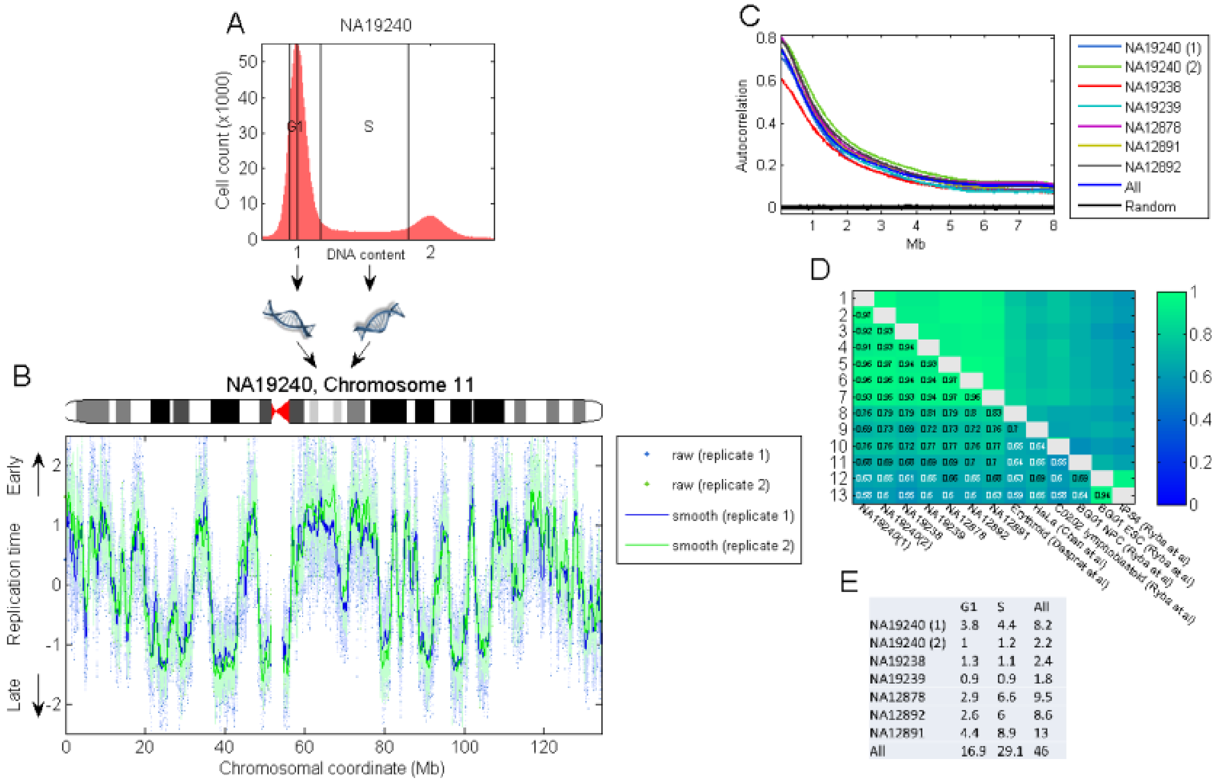
American Journal of Human Genetics, Volume *91*

Supplemental Information

Differential Relationship of DNA Replication Timing to Different Forms of Human Mutation and Variation

Amnon Koren, Paz Polak, James Nemesh, Jacob J. Michaelson, Jonathan Sebat, Shamil R. Sunyaev, and Steven A. McCarroll



	NA19240(1)	NA19240(2)	NA19238	NA19239	NA12878	NA12892	NA12891	Erythroid (Desprat et al)	HeLa (Chen et al)	C0202 lymphoblastoid (Ryba et al)	BG01 NPC (Ryba et al)	BG01 ESC (Ryba et al)	iPS4 (Ryba et al)
1													
2	0.97												
3	0.92	0.93											
4	0.91	0.93	0.94										
5	0.96	0.97	0.94	0.93									
6	0.96	0.96	0.94	0.94	0.97								
7	0.93	0.96	0.93	0.94	0.97	0.96							
8	0.76	0.79	0.79	0.81	0.79	0.8	0.83						
9	0.69	0.73	0.69	0.72	0.73	0.72	0.76	0.7					
10	0.76	0.76	0.72	0.77	0.77	0.76	0.77	0.65	0.64				
11	0.68	0.69	0.68	0.69	0.69	0.7	0.7	0.64	0.66	0.56			
12	0.63	0.65	0.61	0.66	0.66	0.65	0.68	0.63	0.69	0.6	0.69		
13	0.58	0.6	0.55	0.6	0.6	0.6	0.63	0.59	0.65	0.58	0.64	0.96	

	G1	S	All
NA19240 (1)	3.8	4.4	8.2
NA19240 (2)	1	1.2	2.2
NA19238	1.3	1.1	2.4
NA19239	0.9	0.9	1.8
NA12878	2.9	6.6	9.5
NA12892	2.6	6	8.6
NA12891	4.4	8.9	13
All	16.9	29.1	46

Figure S1. Experimental method.

A. Lymphoblastoid cell line DNA is stained, S- and G1-phase cells are FACS-sorted, and DNA extracted and sequenced. The replication profile is obtained from read depth along the chromosomes in S vs G1 cells. We used lymphoblastoid cell lines previously sequenced as part of the 1000 Genomes pilot project (The 1000 Genomes project consortium, 2010). These cell lines are derived from mother-father-offspring trios, one of an African origin (YRI) and the other of a European origin (CEU).

B. Raw (dots) and smoothed (lines) data from two experiment repetitions (blue and green). Replication timing data is normalized to 0 mean and 1 std.

C. Autocorrelation. Black: 100 autocorrelation plots of randomized datasets.

D. Correlation matrix. External data are from: Desprat et al., 2009; Chen et al., 2010; Ryba et al., 2010.

E. Coverage summary.

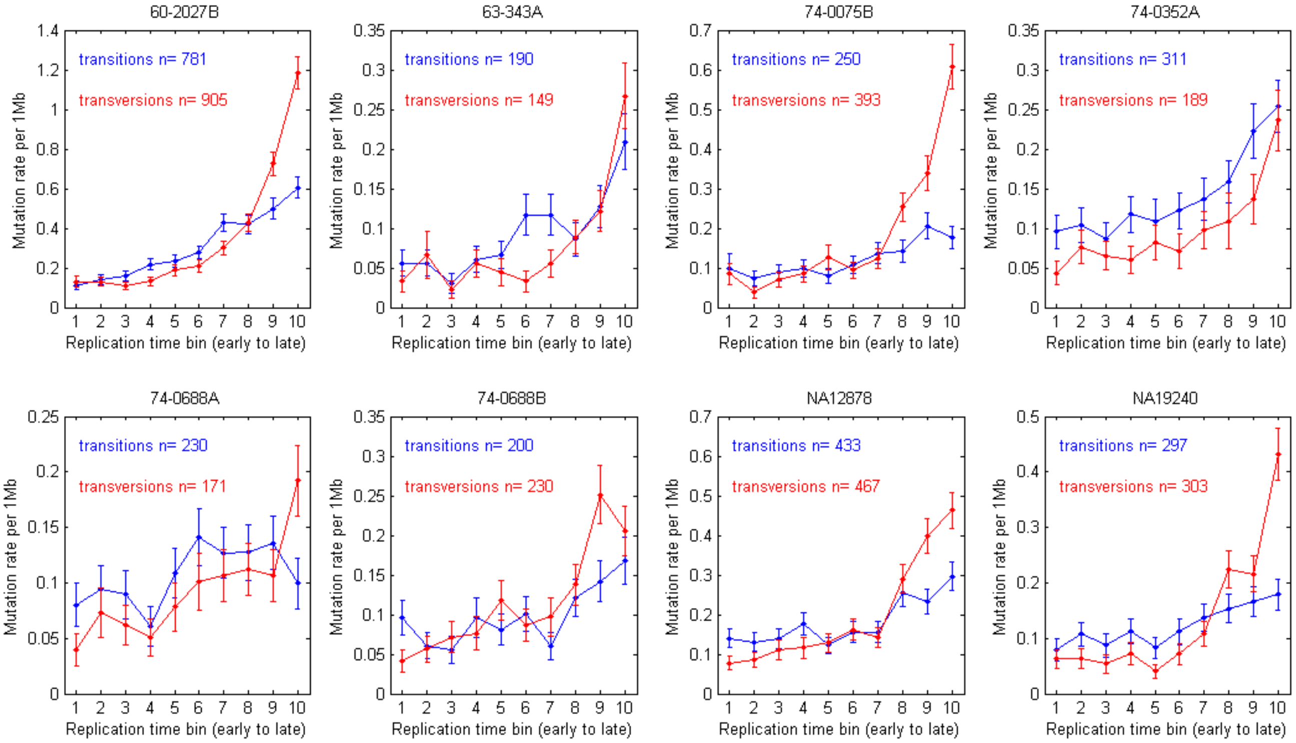


Figure S2. Mutation rate per cell line in 10 replication time bins.
The genome was separated into 10 equally sized bins of increasing replication timing (from early to late). In each bin, the rate of transition and transversion mutations was calculated in 1Mb-sized windows amd the mean and standard error of all windows in each time bin are shown. This was performed separately for each cell line. Note that most data from the quartet dataset originates from six cell lines (the different number of mutations in each cell line probably reflects the age of the cell lines). Only cell lines with at least 100 transition and transversion mutations are shown. The CEU trio cell line is NA12878 and the YRI trio cell line is NA19240. Different cell lines show consistent continuous increases in mutation rate along the S phase, with a sharper increase in the rate of transversions.

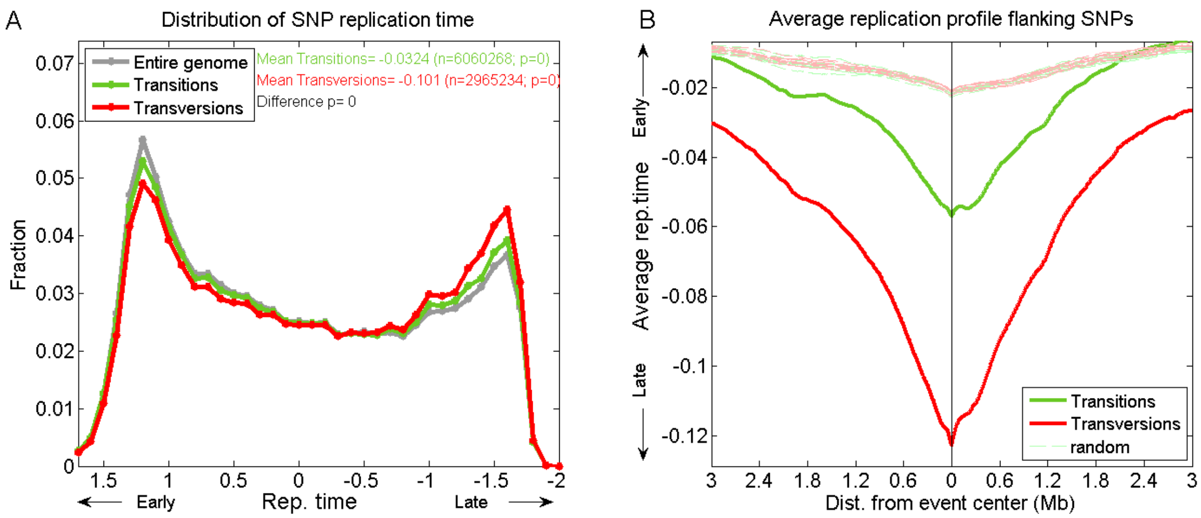


Figure S3. SNP replication time.

A. Distribution of DNA replication timing for the entire genome (grey; mean set to 0) and for SNPs from the 1000 genomes pilot project CEU panel (green: transitions; red: transversions).
B. The average replication timing structure in the region extending to 3 megabases of both sides of all SNP locations. In dashed lines are the same plots for 20 sets each of random genomic locations matching in number to the mutation events of the different types. The pattern of the random tracks in this case results from: 1) the downward pattern is the result of the structure of the replication profile (long late domains). 2) the negative sign is the result of not including regions proximal to gaps (only regions without gaps in the flanking 3Mb were included). The pattern is seen here (but not in other similar figures analyzing different datasets) because of the large number of events (1 million random locations per track)

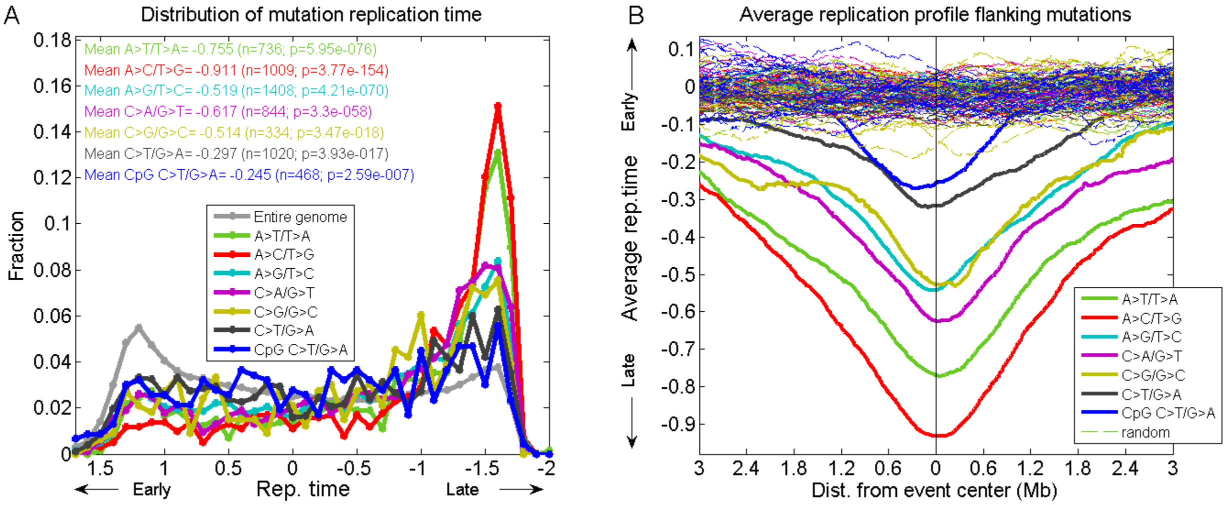


Figure S4. Replication time of different cell line mutation types.

A. Distribution of DNA replication timing for the entire genome (grey; mean set to 0) and for the different types of nucleotide substitution mutation.
B. The average replication timing structure in the region surrounding all mutation locations, and 20 matched sets of random locations. See Figure S3B legend for more details.

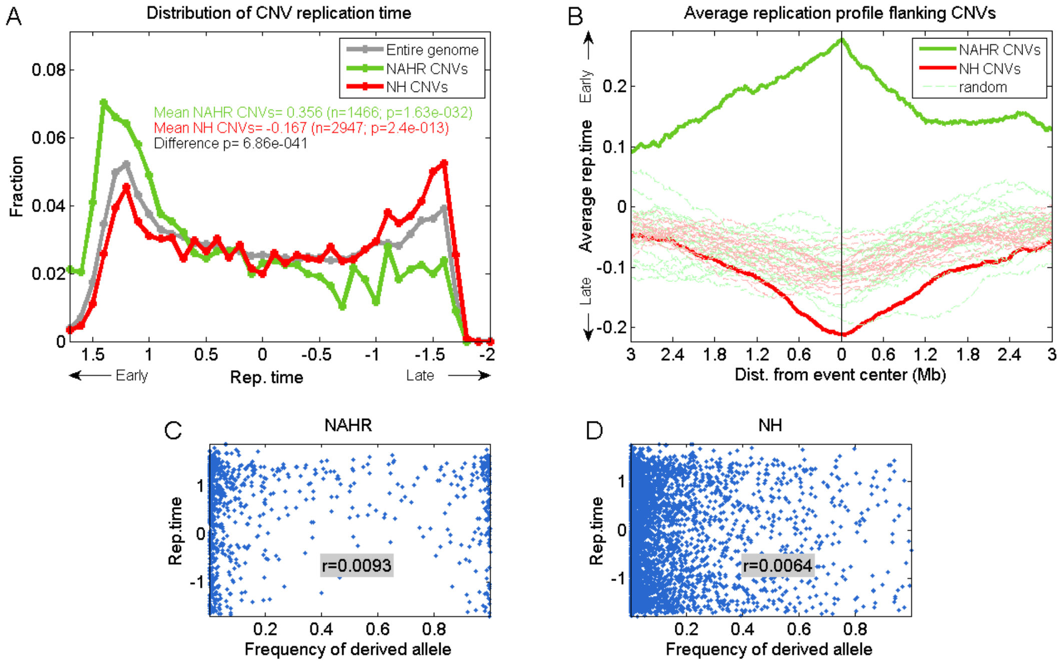


Figure S5. Controlling for functional and repetitive elements and for selection.

A. Distribution of DNA replication timing for the entire genome (grey) and locations of all NAHR (green) and NH (red)-mediated CNVs, excluding CNVs overlapping regions that contain functional elements (genes, CpG islands, lincRNAs, conserved noncoding sequences) or segmental duplications. Consistent results were obtained when removing events whose breakpoints overlap L1 or Alu mobile element sequences (not shown). The deviations of the CNVs from the genome average are comparable to those obtained when looking at all CNV events (mean NAHR=0.388, n=2254; mean NH=-0.118, n=5167).
B. The average replication timing structure surrounding CNV locations (as in Figure S3B) that do not overlap regions with functional elements (genes, CpG islands, lincRNAs, conserved noncoding sequences) or segmental duplications.
C. Allele frequency does not correlate with DNA replication timing. Shown is CNV replication timing versus the frequency of the derived allele (deletion or duplication) for NAHR (C) and NH (D); fixed alleles were removed from this analysis.

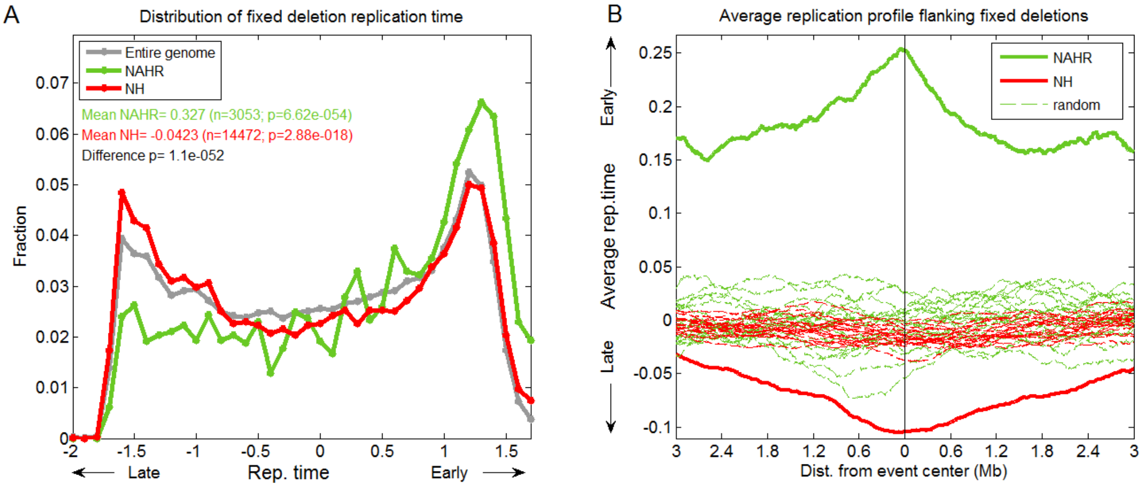


Figure S6. Replication time of fixed human deletion loci.

A. Distribution of DNA replication timing for the entire genome (grey), locations of human-specific deletions putatively mediated by NAHR (green) or by NH (red). Data from human-specific deletions is from McClean et al., 2011; deletions with >50bp with >70% identity flanking the breakpoints were classified as NAHR, and those with <70% identity as NH.
B. The average replication timing structure surrounding fixed deletion locations (as in Figure S3B).

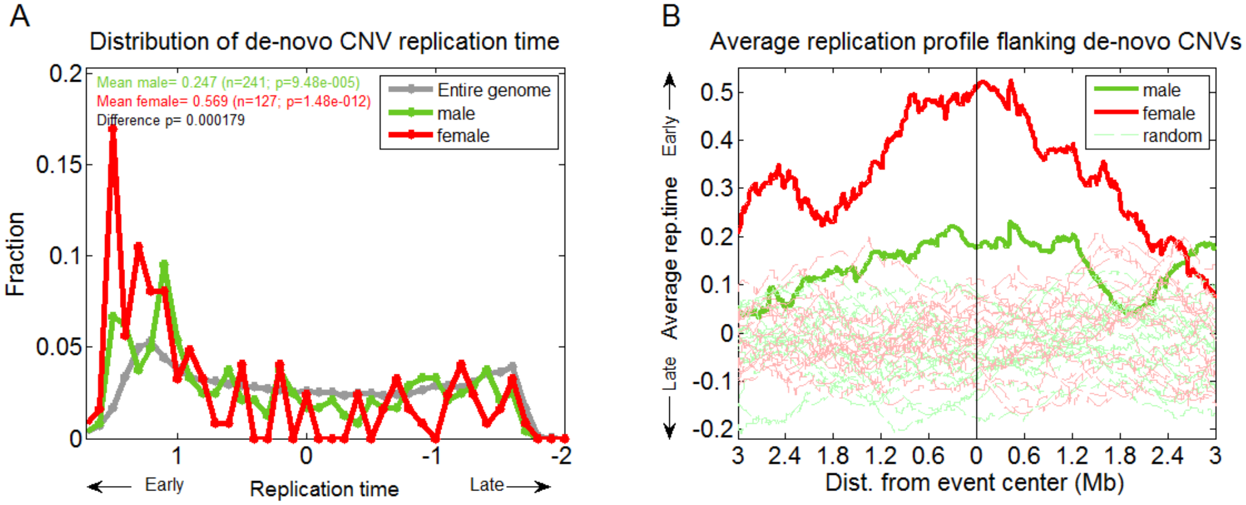


Figure S7. Sex-specific associations of recombination hotspots and *de-novo* CNVs with DNA replication timing.

A. Distribution of DNA replication timing for the entire genome (grey) and locations of male (green) and female (red) *de-novo* CNVs. Data based on SNP array experiments for family trios, from: Kirov et al., 2011; Hehir-Kwa et al., 2011; Itsara et al., 2010; Sanders et al., 2011; Sibbons et al., 2012.
B. The average replication timing structure surrounding *de-novo* CNVs.

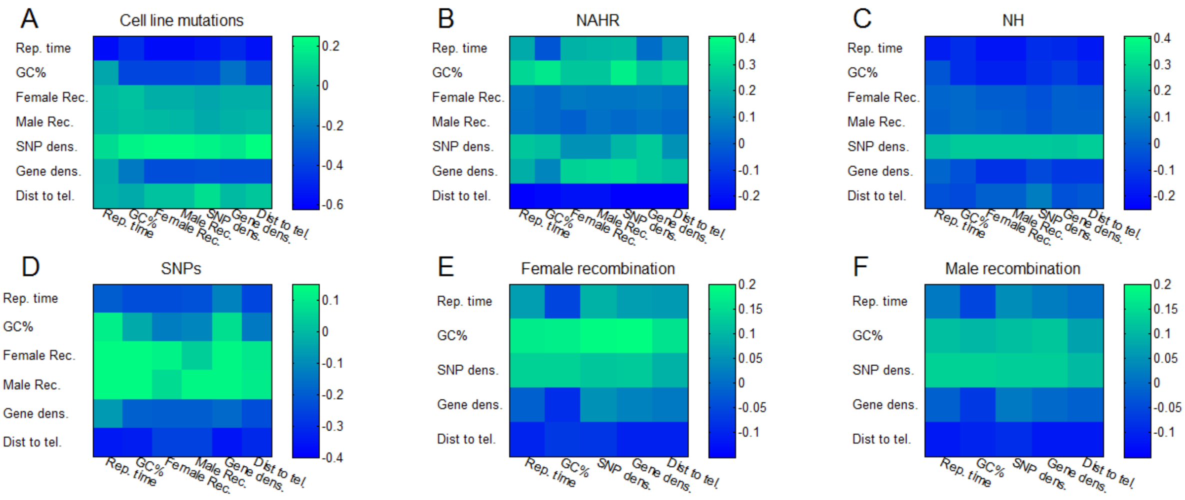


Figure S8. Partial correlations.
Each matrix shows the correlation and partial correlation of one type of genetic variant with the indicated factors. The density of the different variation types was calculated in windows of 1Mb. Predictors were binned in 100Kb windows. The diagonal shows the (complete) pearson correlation of the genetic variation type with the respective factor indicated in either axes; all other data points show the partial correlation between the genetic variation type and factor indicated in the y-axis, controlling for the factor on the x-axis. Genetic variation types are: A) Cell line mutations. B) NAHR CNVs. C) NH CNVs. D) SNPs. E) Female recombination. F) Male recombination. Note scale differences between panels.
The main conclusions from this analysis are: the positive correlation of NAHR with replication timing is lost when controlling for GC content, while the correlation of NAHR with GC content is stronger and robust to the effects of replication timing (and the other factors); In contrast, the negative correlation of NH with replication timing is robust to effects of covariates; The positive correlation of female recombination hotspots with replication timing is also confounded by GC content; on the other hand, when controlling for GC content, male recombination rates show a negative correlation with replication timing; mutations show a strong and genuine negative correlation with replication timing, and replication timing also confounds the negative correlation of mutation rate with GC content; SNPs are associated with replication timing (in a negative direction) as well as with recombination and distance to the telomere.

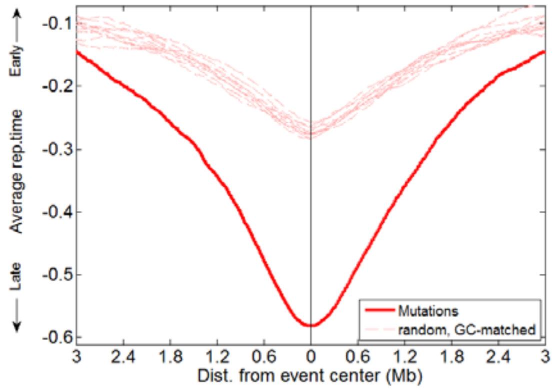


Figure S9. Controlling for GC content effects.
The average replication timing structure surrounding mutations, together with a set of 10 randomized sets of genomic locations matched to have a similar (within 0.01%) GC content. The associations of cell line mutations with DNA replication timing is not due to GC effects,

Supplemental references

Desprat, R., D. Thierry-Mieg, et al. (2009). "Predictable dynamic program of timing of DNA replication in human cells." Genome Research 19(12): 2288-2299.

Hehir-Kwa, J. Y., B. Rodriguez-Santiago, et al. (2011). "De novo copy number variants associated with intellectual disability have a paternal origin and age bias." Journal of Medical Genetics 48(11): 776-778.

Itsara, A., H. Wu, et al. (2010). "De novo rates and selection of large copy number variation." Genome Research 20(11): 1469-1481.

Kirov, G., A. J. Pocklington, et al. (2012). "De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia." Molecular Psychiatry 17(2): 142-153.

Sanders, S., A. J., A. G. Ercan-Sencicek, et al. (2011). "Multiple Recurrent De Novo CNVs, Including Duplications of the 7q11.23 Williams Syndrome Region, Are Strongly Associated with Autism." Neuron 70(5): 863-885.

Sibbons, C., J. K. Morris, et al. (2012). "De novo deletions and duplications detected by array CGH: a study of parental origin in relation to mechanisms of formation and size of imbalance." European Journal of Human Genetics 20(2): 155-160.

Table S1. Poison regression results.

		Estimate	Std. Error	z value	Probability (>|z|)
Transitions	(Intercept)	-13.4554	0.708394	-18.9942	1.90E-80
	GC content	-0.30907	0.802288	-0.38523	0.700065
	SNP density	0.00044	0.000104	4.223422	2.41E-05
	Replication time	-0.45397	0.038635	-11.7502	7.05E-32
	Male_rec	-0.00329	0.023983	-0.13725	0.890835
	Female_rec	0.023754	0.029523	0.804576	0.421065
	LDT	-0.04267	0.031556	-1.35233	0.176269
Transversions	(Intercept)	-11.5117	0.657224	-17.5157	1.09E-68
	GC content	-5.29762	0.82924	-6.38853	1.67E-10
	SNP density	0.000396	0.000112	3.545763	0.000391
	Replication time	-0.77976	0.037242	-20.9376	2.43E-97
	Male_rec	-0.00607	0.021866	-0.27762	0.781308
	Female_rec	0.053461	0.026812	1.993954	0.046157
	LDT	-0.03597	0.028034	-1.28295	0.199511
SNPs	(Intercept)	-4.11392	0.008759	-469.703	0
	GC content	0.095043	0.00939	10.12177	4.42E-24
	Replication time	-0.08732	0.000478	-182.688	0
	Male_rec	0.01254	0.00029	43.30262	0
	Female_rec	0.057634	0.000348	165.7865	0
	LDT	-0.07462	0.000402	-185.743	0

NAHR	(Intercept)	-14.8598	1.120958	-13.2563	4.15E-40
	GC content	6.748416	1.173133	5.752473	8.79E-09
	SNP density	0.000808	7.06E-05	11.4482	2.40E-30
	Replication time	-0.43078	0.063837	-6.74811	1.50E-11
	Male_rec	-0.06179	0.040714	-1.51776	0.129075
	Female_rec	0.032452	0.045831	0.708066	0.478904
	LDT	-0.19451	0.051402	-3.78401	1.54E-04
NH	(Intercept)	-12.7577	0.466142	-27.3688	6.45E-165
	GC content	-1.51673	0.512528	-2.95932	0.003083
	SNP density	0.000672	3.65E-05	18.41353	1.02E-75
	Replication time	-0.12468	0.024801	-5.02722	4.98E-07
	Male_rec	-0.00775	0.016239	-0.4774	0.633077
	Female_rec	0.02958	0.01936	1.527926	0.126531
	LDT	-0.01226	0.021113	-0.58062	5.61E-01
Female recombination	(Intercept)	-18.7034	0.389693	-47.9952	0
	GC content	7.639812	0.406098	18.81273	5.94E-79
	SNP density	0.000813	2.59E-05	31.3639	6.29E-216
	Replication time	-0.06531	0.025505	-2.56057	0.01045
	LDT	0.099793	0.016112	6.193848	5.87E-10
Male recombination	(Intercept)	-12.4693	0.346636	-35.9723	2.27E-283
	GC content	3.197879	0.398739	8.019989	1.06E-15

	SNP density	0.000792	2.72E-05	29.08354	5.80E-186
	Replication time	-0.19013	0.023208	-8.19278	2.55E-16
	LDT	-0.14625	0.013706	-10.6699	1.41E-26

LDT: log distance to telomere.